dbGaP Study Release Notes

Release Notes for NHLBI TOPMed WGS SAGE, phs000921.v4.p1
"NHLBI TOPMed: Study of African Americans, Asthma, Genes and Environment (SAGE)"

For any questions or comments, please contact: dbgap-help@ncbi.nlm.nih.gov.

October	17, 2016	Version 1 Data set release date
February	7, 2018	Version 2 Data set release date
June	19, 2018	Version 3 Data set release date
January	6, 2020	Version 4 Data set release date

2020-01-06
Version 4 Data set release for NHLBI TOPMed WGS SAGE now available

This release includes the addition of Freeze 8 whole genome sequences (WGS) brokered through the Sequence Read Archive (SRA), and VCFs derived from WGS. Please refer to the latest study configuration report for a detailed description of each download component.

Consent name and Data Use Limitation (DUL) updated

Consent group 2 (c2) is changed from Disease-Specific (Lung Disease) (DS-LD) to Disease-Specific (Lung Diseases, IRB, COL) (DS-LD-IRB-COL).

The following is the current Data Use Limitation:

Use of the data must be related to Lung Diseases.
Requestor must provide documentation of local IRB approval.
Requestor must provide a letter of collaboration with the primary study investigator(s).

Consent group 2 (c2): Disease-Specific (Lung Diseases, IRB, COL) (DS-LD-IRB-COL)

Data Type	subjects	samples
Phenotype	2106	2105
Seq_DNA_SNP_CNV (VCFs)	1840	1840
WGS*	1840	1840

*These data are brokered through the Sequence Read Archive (SRA). Please see Authorized Access instructions below.
For a description of non-SRA SAMPLE_USE terms, please see:
https://www.ncbi.nlm.nih.gov/projects/gap/submission/GetSampleUseTypes.cgi

Study and Phenotype Data Updates

1. **New Study Accession**

   NHLBI TOPMed WGS SAGE version 3 phs000921.v3.p1 has been updated to Version 4. The dbGaP accession for the current set of data is **phs000921.v4.p1**. The participant number (p#) has not changed in version 4. No new subjects have been added to this study.

2. **Updated Datasets (n=1 dataset)**

pht	version	Dataset Name

4881	4	TOPMed_WGS_SAGE_Subject

3. **Retired Variables (n=2 variables)**

pht	Dataset Name	phv	version	Variable Name
4881	TOPMed_WGS_SAGE_Subject	252276	3	SUBJECT_SOURCE
4881	TOPMed_WGS_SAGE_Subject	252277	3	SOURCE_SUBJECT_ID

Molecular Data Updates

dbGaP QC steps for this release consist of checks for consistency of subject and sample IDs in phenotype and genotype components.

1. For samples and marker/enrichment-procedure info, see download components:
    a. phg001074.v2.TOPMed_WGS_SAGE_v4.sample-info.MULTI.tar.gz
    b. phg001074.v2.TOPMed_WGS_SAGE_v4.marker-info.MULTI.tar.gz
    c. phg001316.v1.TOPMed_WGS_SAGE_v4.sample-info.MULTI.tar.gz
    d. phg001316.v1.TOPMed_WGS_SAGE_v4.marker-info.MULTI.tar.gz
2. Genotypes are available in a matrix format as multi-sample vcf file(s) packed within download component(s) marked as genotype-calls-vcf. Integrity of submitted vcf files and their compatibility with PSEQ are routinely checked. Components may be divided by platform and/or population.
    a. phg001074.v2.TOPMed_WGS_SAGE_v4.genotype-calls-vcf.WGS_markerset_grc38.c2.DS-HCR-IRB.tar.gz
    b. phg001316.v1.TOPMed_WGS_SAGE_v4.genotype-calls-vcf.WGS_markerset_grc38.c2.DS-LD-IRB-COL.tar.gz

Authorized Access (Individual Level Data and SRA Data)

Individual level data and Sequence Read Archive (SRA) data are available for download through the dbGaP Authorized Access System upon approval of the Data Access Request (DAR):

- http://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?login=&page=login

Public FTP site (Summary Level Data Only)

All data tables, data dictionaries, and documents will be housed under one directory for ease of downloading. The data_dict filenames have an added study version number (phs#.v#) and deleted participant set number (p#) from the table accession (pht#.v#). The var_report filenames have an added study version number (phs#.v#). In the var_report files, variables contain version numbers (phv#.v#) and summaries were created for each consent group (c#). These FTP files are available at:

- ftp://ftp.ncbi.nlm.nih.gov/dbgap/studies/phs000921/phs000921.v4.p1

2018-06-19

Version 3 Data set release for NHLBI TOPMed WGS SAGE now available

This release includes a second genotype call set (GRCh38) and updated phenotype tables. Please refer to the latest study configuration report for a detailed description of each download component.

Consent group 2 (c2): Disease-Specific (Lung Disease) (DS-LD)

	Phenotype	VCFs	WGS

subjects	2106	499	499
samples	2105	499	499

Study and Phenotype Data Updates

1. **New Study Accession**

   NHLBI TOPMed WGS SAGE version 2 phs000921.v2.p1 has been updated to Version 3. The dbGaP accession for the current set of data is **phs000921.v3.p1**. The participant number (p#) has not changed in version 3. New subjects have been added to this study.

2. **Updated Datasets (n=4 datasets; all existing variables have been updated)**

pht	version	Dataset Name
4881	3	TOPMed_WGS_SAGE_Subject
4882	3	TOPMed_WGS_SAGE_Sample
4883	3	TOPMed_WGS_SAGE_Subject_Phenotypes
4884	3	TOPMed_WGS_SAGE_Sample_Attributes

3. Please note we are discontinuing the submission and distribution of the SAMPLE_USE variable. The sample use counts will be populated by SRA (sequences) and dbGaP (all other submitted molecular data).

Molecular Data Updates

dbGaP QC steps for this release consist of checks for consistency of subject and sample IDs in phenotype and genotype components.

1. For samples and marker/enrichment-procedure info, see download components:
   a. phg001074.v1.TOPMed_WGS_SAGE_v3.sample-info.MULTI.tar.gz
   b. phg001074.v1.TOPMed_WGS_SAGE_v3.marker-info.MULTI.tar.gz
2. The standard dbGaP QC pipeline was applied on SNP genotypes in PLINK format. Results are in tar files marked as "genotype-qc"
3. Genotypes are available in a matrix format as multi-sample .vcf file(s) packed within download component(s) marked as "genotype-calls-vcf. Integrity of submitted .vcf files and their compatibility with PSEQ are routinely checked. Components may be divided by platform and/or population.
   a. phg001074.v1.TOPMed_WGS_SAGE_v3.genotype-calls-vcf.WGS_markerset_grc38.c2.DS-LD.tar.gz

Authorized Access (Individual Level Data and SRA Data)

Individual level data and Sequence Read Archive (SRA) data are available for download through the dbGaP Authorized Access System upon approval of the Data Access Request (DAR):

- http://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?login=&page=login

Public FTP site (Summary Level Data Only)

All data tables, data dictionaries, and documents will be housed under one directory for ease of downloading. The data_dict filenames have an added study version number (phs#.v#) and deleted participant set number (p#) from the table accession (pht#.v#). The var_report filenames have an added study version number (phs#.v#). In the var_report files, variables contain version numbers

(phv#.v#) and summaries were created for each consent group (c#). These FTP files are available at:

- ftp://ftp.ncbi.nlm.nih.gov/dbgap/studies/phs000921/phs000921.v3.p1

2018-02-07

Version 2 Data set release for NHLBI TOPMed WGS SAGE now available

This release includes updated phenotype tables, whole genome sequences (WGS) brokered through the SRA, and VCFs derived from WGS. Please refer to the latest study configuration report for a detailed description of each download component.

Consent group 2 (c2): Disease-Specific (Lung Disease) (DS-LD)

	Phenotype	Seq_DNA_SNP_CNV	Seq_DNA_WholeGenome
subjects	500	499	499
samples	499	499	499

Molecular data descriptions:
(http://www.ncbi.nlm.nih.gov/projects/gap/submission/GetSampleUseTypes.cgi)

a. Seq_DNA_SNP_CNV: SNP and CNV genotypes derived from sequence data (VCFs)
b. Seq_DNA_WholeGenome: Whole genome sequencing

Study and Phenotype Data Updates

1. **New Study Accession**

NHLBI TOPMed WGS SAGE version 1 phs000921.v1.p1 has been updated to Version 2. The dbGaP accession for the current set of data is **phs000921.v2.p1**. The participant number (p#) has not changed in version 2. No new subjects have been added to this study.

2. **Updated Datasets (n=4 datasets; all existing variables have been updated)**

pht	version	Dataset Name
4881	2	TOPMed_WGS_SAGE_Subject
4882	2	TOPMed_WGS_SAGE_Sample
4883	2	TOPMed_WGS_SAGE_Subject_Phenotypes
4884	2	TOPMed_WGS_SAGE_Sample_Attributes

3. **New Variables (n=10 variables)**

pht	pht version	Dataset Name	phv	Variable Name
4883	2	TOPMed_WGS_SAGE_Subject_Phenotypes	347787	smoke_current
4883	2	TOPMed_WGS_SAGE_Subject_Phenotypes	347788	asthma_hospital_12mo
4883	2	TOPMed_WGS_SAGE_Subject_Phenotypes	347789	steroids_12mo
4883	2	TOPMed_WGS_SAGE_Subject_Phenotypes	347790	delta1
4883	2	TOPMed_WGS_SAGE_Subject_Phenotypes	347791	no2_lifetime
4883	2	TOPMed_WGS_SAGE_Subject_Phenotypes	347792	pm25_lifetime
4884	2	TOPMed_WGS_SAGE_Sample_Attributes	347793	Funding Source

4884	2	TOPMed_WGS_SAGE_Sample_Attributes	347794	TOPMed_Phase
4884	2	TOPMed_WGS_SAGE_Sample_Attributes	347795	TOPMed_Project
4884	2	TOPMed_WGS_SAGE_Sample_Attributes	347796	Study_Name

Molecular Data Updates

dbGaP QC steps for this release consist of checks for consistency of subject and sample IDs in phenotype and genotype components.

1. For samples and marker/enrichment-procedure info, see download components:
    a. phg001016.v1.TOPMed_WGS_SAGE_v2.sample-info.MULTI.tar.gz
    b. phg001016.v1.TOPMed_WGS_SAGE_v2.marker-info.MULTI.tar.gz
2. The standard dbGaP QC pipeline was applied on SNP genotypes in PLINK format. Results are in tar files marked as "genotype-qc".
3. Genotypes are available in a matrix format as multi-sample .vcf file(s) packed within download component(s) marked as "genotype-calls-vcf. Integrity of submitted .vcf files and their compatibility with PSEQ are routinely checked. Components may be divided by platform and/or population.
    a. phg001016.v1.TOPMed_WGS_SAGE_v2.genotype-calls-vcf.WGS_markerset_grc37.c2.DS-LD.tar.gz.

Authorized Access (Individual Level Data and SRA Data)

Individual level data and Sequence Read Archive (SRA) data are available for download through the dbGaP Authorized Access System upon approval of the Data Access Request (DAR):

- http://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?login=&page=login

Public FTP site (Summary Level Data Only)

All data tables, data dictionaries, and documents will be housed under one directory for ease of downloading. The data_dict filenames have an added study version number (phs#.v#) and deleted participant set number (p#) from the table accession (pht#.v#). The var_report filenames have an added study version number (phs#.v#). In the var_report files, variables contain version numbers (phv#.v#) and summaries were created for each consent group (c#). These FTP files are available at:

- ftp://ftp.ncbi.nlm.nih.gov/dbgap/studies/phs000921/phs000921.v2.p1

2016-10-17

Version 1 Data set release for NHLBI TOPMed WGS SAGE now available

This release includes TOPMed Phase I phenotype tables, whole genome sequences (WGS) brokered through the SRA, and VCFs derived from WGS. Additionally, phenotype tables include subjects and samples beyond TOPMed Phase I in order to instantiate IDs for future versions. Please refer to the latest study configuration report for a detailed description of each download component.

Consent group 2 (c2): Disease-Specific (Lung Disease) (DS-LD)

	phenotype	SRA/VCFs
subjects	500	485
samples	500	485

Molecular Data Updates

dbGaP QC steps for this release consisted of checks for consistency of subject and sample IDs in phenotype and genotype components:

1. For samples and marker/enrichment-procedure info see download components:
    a. phg000790.v1.TOPMed_WGS_SAGE.sample-info.MULTI.tar.gz
    b. phg000790.v1.TOPMed_WGS_SAGE.marker-info.MULTI.tar.gz
2. Genotypes are available in a matrix format as multi-sample .vcf file(s) packed within download component(s) marked as "genotype-calls-vcf. Integrity of submitted .vcf files and their compatibility with PSEQ are routinely checked. It is noted when components are divided by platform and/or population.
    a. phg000790.v1.TOPMed_WGS_SAGE.genotype-calls-vcf.WGS_markerset_grc37.c2.DS-LD.tar.gz
3. The standard dbGaP QC pipeline was applied on SNP genotypes in PLINK format. Results are in tar files marked "genotype-qc".

Authorized Access (Individual Level Data and SRA Data)

Individual level data and SRA sequencing data are available for download through the dbGaP Authorized Access System upon approval of the Data Access Request (DAR):

- http://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?login=&page=login

Public FTP site (Summary Level Data Only)

All data tables, data dictionaries, and documents will be housed under one directory for ease of downloading. The data_dict filenames have an added study version number (phs#.v#) and deleted participant set number (p#) from the table accession (pht#.v#). The var_report filenames have an added study version number (phs#.v#). In the var_report files, variables contain version numbers (phv#.v#) and summaries were created for each consent group (c#). These FTP files are available at:

- ftp://ftp.ncbi.nlm.nih.gov/dbgap/studies/phs000921/phs000921.v1.p1